DOCUMENT RESUME

ED 475 161 SE 067 680

AUTHOR	Schneider, Rebecca; Blumenfeld, Phyllis
TITLE	Observing Teaching: A Reform-Based Framework for Looking into Classrooms.
SPONS AGENCY	National Science Foundation, Arlington, VA.
PUB DATE	2003-03-00
NOTE	30p.; Produced by the Center for Learning Technologies in Education.
CONTRACT	NSF-0830-310-A605
AVAILABLE FROM	Web site: http://hi- ce.org/teacherworkroom/middleschool/ physics/index.html.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Classroom Research; Curriculum Development; Educational Change; *Lesson Observation Criteria; Middle Schools; Professional Development; *Science Education; *Science Teachers; Student Projects; Teaching Methods

ABSTRACT

This study examines teachers' classroom practices in response to the support for teacher thinking in the materials and designs a systematic research method for observing classroom teaching consistent with reform recommendations and adaptable for use on a large scale. The development of a method to evaluate complex classroom observations that captures the salient features of reform-based teaching is described. (KHR)

Running Head: Observing teaching

Observing Teaching: A reform-based framework for looking into classrooms

Rebecca Schneider

University of Toledo

Phyllis Blumenfeld

University of Michigan

National Association for Research in Science Teaching
Philadelphia 2003

This study was funded in part by the National Science Foundation as part of the Center for Learning Technologies in Education grant 0830 310 A605. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

More information about this work including the curriculum materials "Why do I need to wear a bike helmet?" used in this study, can be obtained from our project's web site at this address: http://hi-ce.org/teacherworkroom/middleschool/physics/index.html

Correspondence concerning this article should be addressed to Rebecca Schneider, College of Education, University of Toledo, Toledo, Ohio 43606. E-mail: Rebecca.Schneider@utoledo.edu

Observing Teaching: A reform-based framework for looking into classrooms

Introduction

Teaching and its measure are critical components of efforts to promote student learning (National Commission on Teaching and America's Future, 1996). Reformers exploring ways to support exemplary teaching and thus advance student learning, depend on suitable methods for gathering information on what is effective. In our work we are exploring the use of reform-based curriculum materials to promote exemplary teaching in science. We have designed explicit support for teachers to learn about teaching within our materials (Schneider & Krajcik, 2002). This work has led us to examine teachers' classroom practices in response to the support for teacher thinking in the materials. Observation of classroom teaching is essential to improve our understanding of how to help teachers learn and enact reform-based practices (Anderson, 2001). However, the rich descriptions provided by qualitative methods are time and labor intensive necessitating the observation of only a few teachers. Data from a variety of classrooms is needed to develop truly effective programs. We also are interested in the scalability of our teacher educative materials and therefore a measure of teaching that is less cumbersome than detailed descriptions of classroom events. Although many quantitative measures are feasible on a large scale they fail to capture the true complexity of what happens in classrooms. In this paper we describe the development of a method to evaluate complex classroom observations that captures the salient features of reform-based teaching and is feasible on a larger scale.

Observing teaching

Based on goals for student learning in science, reformers are exploring new ways to help teachers learn how to use inquiry with collaboration supported by use of technology tools to support students in actively constructing deep understanding of important science concepts

(National Research Council, 1996). One new idea is to include explicit support for teachers to learn about teaching within curriculum materials making them educative for teachers (Ball & Cohen, 1996). This strategy has the potential to facilitate instructional improvement on a large scale but lacks specific design ideas and empirical evidence that it can be effective. Research in classrooms is needed to guide the design and improvement of educative materials.

We have developed science materials to reflect desired reforms and provide teachers with needed support to learn and enact innovative curriculum as part of an ongoing systemic initiative of a large urban public school district. Developers created materials based on the premises of project-based science and were guided by design principles that include: contextualization, alignment with standards, sustained student inquiry, embedded learning technologies, collaboration and discourse, assessment techniques, and scaffolds and supports for teachers (Krajcik, Czerniak, & Berger, 2002; Schneider & Krajcik, 2002; Singer, Marx, Krajcik, & Clay-Chambers, 2000). Materials were designed to be educative by including detailed lesson descriptions that addressed necessary content, pedagogy, and pedagogical content knowledge for teachers. We were interested in a scalable method of analyzing classroom enactment data to gain meaningful information on which to base revisions of curriculum materials and improve support for teachers in learning and enacting new instructional practices.

Researchers interested in understanding how to support improved teaching consider classroom observation essential to determining the success of their efforts to change teachers' practice (Blumenfeld, Krajcik, Marx, & Soloway, 1994; Palincsar, Magnusson, Marano, Ford, & Brown, 1998; Wood, Cobb, & Yackel, 1991). This requires careful observation and analysis of classroom events including teachers' behaviors and statements. Typically this process extends to only one to four classrooms over a period of one to four years. Therefore, we have only

speculative knowledge as to why many wonderful curriculum ideas fail to be realized in classrooms beyond the initial implementations (Brown, 1992).

Researchers attempting to identify specific factors that influence student achievement are examining national and state level data. From this work we have some evidence that the quality of teaching is related to student outcomes. For instance, Darling-Hammond (1999) examined state level data on teacher preparation, certification, and experience along with changes in student achievement over several years. She describes teacher professional development as the most important means to improving student achievement scores. However, this approach does not identify what these teachers are doing in the classroom to impact student learning. Likewise, work by Sanders and Horn (1998) indicates a long term affect of individual teachers on student achievement scores. But again this work does not describe what this quality teaching looks like in a classroom. Therefore, these studies cannot point to the features of teacher preparation, knowledge, or experience that are particularly worthwhile. This leaves the topic of how to improve teaching and student outcomes open to debate (Cochran-Smith & Fries, 2001).

One approach that merits further development is classroom observation research that links specific curriculum to teachers' instruction (Collopy, 1999; Prawat, 1992; Remillard, 1999). In these studies, analysis of observations is guided by frameworks based on recommended curriculum. This approach is facilitated when researchers describe their curriculum in terms of reform guidelines. The reform-based curriculum that is the focus of this study is one such example (see Schneider & Krajcik, 2002; Singer et al., 2000).

An better understanding of how teachers and students interact around specific materials and ideas in classrooms is needed (Ball, 2000). Teaching is a thinking practice; it is not enough to measure teachers' knowledge or behaviors independently (Lampert, 1998). We need to look

into classrooms to observe teachers' practices in light support for teacher thinking. Similarly, measures of student achievement matched to specific curriculum are more likely to capture the impact on student achievement than general measures. (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). General measures of student achievement do not necessarily indicate what students have learned in the classroom. In this study we used a reform-based science curriculum unit to develop a framework for observing teaching and linked our observation results to student outcomes also measured closely to the curriculum project.

Research design

The goal of this research was to design a systematic research method for observing classroom teaching that was consistent with reform recommendations and adaptable to use on a large scale. To support teachers in science reform, project-based materials were developed to address important science ideas, offer multiple learning opportunities, and provide appropriate instructional supports for students. The materials also incorporate ideas about how and what teachers need to learn to enact innovative curriculum. Materials include detailed lesson descriptions to assist teachers in enactment. Features to address the learning needs of teachers offer information to explain content and pedagogy, as well as specific information about strategies, representations, and students' ideas (PCK) embedded within lessons (Schneider & Krajcik, 2002). In order to determine if educative materials were indeed helpful for teachers enacting project-based science a careful qualitative study was conducted. The results of that work have been reported in two other papers (Schneider, Blumenfeld, & Krajcik, 2002; Schneider, Krajcik, & Blumenfeld, 2002). Teachers' enactments of these lesson sequences were examined and characterized across lessons and teachers in light of the instructional practices

recommended in the materials. One outcome of this work is the beginning of a scalable scoring rubric to measure quality of teaching in comparison to curriculum goals.

Methods

Background

This study was conducted in four urban middle schools located in low SES neighborhoods selected to participate in initial stages of the reform effort (Krajcik, Marx, Blumenfeld, Soloway, & Fishman, 2000). Students in these schools were predominantly African American (95% to 100%) with high percentages of students receiving free or reduced lunch (29% to 66%). Scores on local and statewide achievement testing in science were reported as below grade level in three of the four schools.

Curriculum material development was considered an essential component of the change effort, particularly to facilitate change within classrooms on a large scale (Blumenfeld, Fishman, Krajcik, Marx, & Soloway, 2000; Singer et al., 2000). The project-based science curriculum materials used by teachers in this study were developed as part of the larger reform effort. As a researcher and curriculum developer, the first author took a lead role in designing these materials to support both students and teachers in the transition to inquiry based science instruction (see Schneider & Krajcik, 2002; Schneider, Krajcik et al., 2002). However, the educative features of the materials were only one part of the professional development involved in this reform effort (Fishman & Best, 2000).

Four eighth-grade teachers participating in the reform effort used materials for a ten-week unit on force and motion. Teaching experience ranged from 6 to 20 years. Prior to enacting this unit, each of the four teachers had limited experience with one or more of the following aspects: project-based science, physics, or the use of technological tools to support inquiry. Although

they were not selected as a statistically random sample, their disparate backgrounds made this group representative of middle school science teachers across the district.

Data Collection and Preparation

Target lesson sequences. Five target lesson sequences containing experiences with phenomena, investigation, technology use, or artifact development, spanning 3-5 days each were selected for analysis. These lesson sequences were selected because each represented different aspects of inquiry teaching that were to be used to focus descriptions of classroom enactments. These aspects included how teachers a) presented science ideas, b) promoted students' use of inquiry, c) used technology to promote student inquiry and concept development, d) used collaboration to promote student inquiry and concept development, and e) supported and assessed concept development through student artifacts.

Materials descriptions. Summary descriptions of the materials were created to guide analysis of classroom enactments. Text relevant to each of the five target lesson sequences was selected. Text was coded for categories relevant to supporting student learning. These included: science ideas, contextualization, representation, strategy, suggested instructional supports, collaboration, and artifact development. The coded text was summarized to describe the intended enactments in terms of the goals for student learning, the opportunities for student learning, and suggested instructional supports for each minor and major learning opportunity identified from the text.

Enactment descriptions. Detailed descriptions of classroom events were written from the videotape for each target lesson sequence and teacher. Teacher and student behavior and conversation were described in light of the lesson sequence descriptions in the materials. As these descriptions were prepared, we looked for and described: 1) science ideas (content and

process ideas presented), 2) contextualization (referring to the driving question or anchor ideas, using real life examples, stating value), 3) linking ideas to previous or future lessons or to other ideas, 4) directions given, 5) emphasis given—such as what ideas or tasks are important, 6) specific strategies such as POE, 7) specific representations such as motion graphs, 8) scaffolding (modeling, coaching, feedback, or asking for justifications or reasons), and 9) group work (teacher statements on group work, teacher role during group work). We also noted suggested lesson sequences or portions of lesson sequences that were enacted, omitted, or adapted. Finally, descriptions of instruction were aligned with the intended opportunities for student learning as identified in the description of the materials and labeled accordingly.

Data Analysis

The coding scheme used was designed to capture three aspects of enactment—presentation of science ideas, opportunities for student learning, and support to enhance the learning opportunities—each in comparison to what was intended in the materials. coding schemes used in this analysis were developed through an iterative process of creating codes, coding, modifying and refining codes, and recoding consistent with Miles and Huberman's (1994) recommendations for rigorous and meaningful qualitative data analysis. The independent coding of several enactment episodes by another science education researcher assessed reliability of the coding process. Reliability was 88%. After the categories and rating levels were finalized and reliability established, all enactment data were recoded with the final codes.

The final coding scheme assessed instructional events in the following groups of categories 1) *accuracy* and *completeness* of the science ideas presented, 2) the amount student learning *opportunities*, *similarity* of learning opportunities with those intended, and quality of the *adaptations,* and 3) the amount of *instructional supports* offered, the *appropriateness* of the

instructional supports and the *source* of ideas for instructional supports. Each enactment episode was rated in each category according to the descriptions listed in Table 1 for each rating level. Entire episodes and the type of activity were considered to assign a rating for each category. A short statement of evidence or justification was written for each assigned rating.

Assigning ratings. The categories of accuracy and completeness were included to capture information about the science ideas presented by teachers. Both content and process ideas were considered as well as whether the ideas presented were defined as a main or minor idea. The main ideas were defined as those identified in the purpose, objectives, or assessments of the materials for that lesson sequence. Likewise, minor ideas were defined as ideas secondary, related, or supporting the main ideas. Teachers presented ideas in a variety of ways. This included teachers' statements, examples, demonstrations, hints, or other types of guidance regarding science ideas. A teacher's response or lack of response to students' actions or statements was also judged as giving students information about science ideas. In this case, a teacher may not have directly stated ideas accurately or inaccurately but, by the type of response they gave, implied that inaccurate student statements were acceptable or vise versa. Each type of presentation of all ideas was considered when rating both accuracy and completeness.

The rating of accuracy was unrelated to the rating of completeness. A rating of scientific for accuracy but incomplete or insufficient for completeness was possible and occurred. Also, unlike any other category, completeness included one rating that could apply in addition to the other ratings. This was the rating of excessive. This rating was used to indicate content related but beyond that intended for students in this unit. A teacher could be incomplete in covering the intended content, yet also excessive by adding other related content. For example, in a lesson on velocity a teacher might not address the intended ideas of speed and direction as components of

motion but might include the formula to calculate speed, which was not intended. This would be rated as incomplete and excessive.

The categories of opportunities, similarity, and adaptation each refer to the learning opportunities for students observed in the episode. Opportunities for student learning included both teacher lead and small-group activities. Take-home activities that were incorporated into class activities were included as opportunities, but work completed entirely at home was not. The number of activities and the amount time devoted to these activities was considered in light of how the enactment episode was segmented. Episodes were not given lower ratings because the enactment was divided into several short segments. Opportunities were rated high if the number and time spent was high in relationship to the amount of class time represented in the episode.

Similarity was rated by considering both that opportunities observed were intended by the materials, but also that they were in a similar sequence with approximately the same emphasis. For example, if a teacher directed students to make a prediction, but did not allow time for writing the predictions or for sharing some of the predictions in class before the observation phase, similarity would be rated low.

Adaptations were opportunities provided that were not described in the materials. These activities were judged on whether or not they addressed content specified for the learning sequence and if the activity was likely to help students learn the content. Replacing a discussion of observed phenomena with a drill and practice to define terms would be rated as low. The terms may be the ones intended for use but understanding of relationships or application of ideas was the intended learning goal rather than the memorization of definitions. On the other hand, making an investigation more open by allowing students more choices in what to test would be rated high if students appeared to be ready to design an investigation with reduced structure.

The categories of instructional supports, appropriateness, and sources each refer to the instructional support for student thinking observed in the episode. Instructional supports included wide variety of teacher actions and statements that had the potential to enhance the learning opportunities. These included supports for student thinking as well as supports for organizing and carrying out tasks. Examples included, but were not necessarily limited to: modeling thought processes or actions, coaching, giving hints, using examples, monitoring small-group work, giving reminders, asking for reasons or justification, structuring student work, offering guidance, and giving feedback.

Instructional supports were rated high if the number of supports was high. Whether or not the supports appeared to be of a type that would help students learn the intended science content was judged in the category of appropriateness. Therefore, an episode could be rated high for supports if a teacher gave students many hints, but poor for appropriateness if those hints were likely to lead students in the wrong direction or did not match the type of difficulty students were exhibiting. The category of source was rated as matched when teachers used only supports that were suggested in the material for that lesson sequence. If the support was suggested in another lesson sequence, source was rated replaced or supplemented. If teachers used only supports not suggested in the materials, source was rated replaced. Supplemented was a rating used when supports of both types were observed.

Summarizing ratings. Ratings were then summarized across opportunities for each lesson sequence. The ratings and the justification statements in each category were compared sequentially for all enactment episodes already rated by opportunity. Then a judgment was made for a rating of the entire lesson sequence. A justification statement was also written for each lesson sequence rating based on a summary of the individual statements. To guide the

summarization process a set of guidelines were developed. When variation was evident, summarizing was done in a way that appropriately reflected the variation in the final rating and justification statement. If the variation was minor, one rating was given but the variation was described in the justification statement. However, when variation was more pronounced, two or more ratings were assigned and the lesson sequence was labeled as varied. Again the justification statement described the variation.

The final analysis phase was to examine the coded lesson sequences for patterns across lesson sequences and teachers. Each category was traced across all lesson sequences for each teacher. During this examination, justifications for the ratings were also examined for patterns. Data also were examined in the same way for patterns across teachers. Summarizing across all teachers was not possible. However, summary ratings and justification statements were appropriate when teachers were placed into two enactment groups.

Student Achievement Measures

As part of the larger research effort in which this study was embedded, written assessment instruments were developed to assess student understanding of the curriculum content and science process skills (Krajcik et al., 2000). The assessments were administered to each student participating in the curriculum projects. The assessments consisted of a combination of multiple choice and free response items that were further classified as either curriculum *content knowledge* or *science process* skill items. Content and process items were categorized by one of three cognitive levels required for arriving at a complete answer: *lower* (recalling information; understanding simple and complex information); *middle* (drawing or understanding simple relationships; applying knowledge to new or different situations; shifting between representations such as verbal to graphic; identifying hypotheses, procedures, results, or

conclusions); and *higher* (describing or analyzing data from charts and graphs; framing hypotheses; drawing conclusions; defining or isolating variables given in a scenario; applying investigation skills; and using concepts to explain phenomena). The curriculum development teams (including science educators, content specialists, educational psychologists, and classroom teachers) constructed the tests. We analyzed all potential questions according to the scheme described above with teams of three to five raters achieving 95% accuracy in categorizing items. Disagreements were settled by consensus. The use of rubrics for each open-ended question produced over 95% agreement by two to four raters each. Again, disagreements were settled by consensus.

Findings

The coding categories and rating levels captured differences in enactment by teacher throughout all lesson sequences. Ratings also indicated teachers were fairly consistent in their enactments. This finding was backed up by the descriptions of specific observation of enactment written in the justification statements. More importantly, this method of describing enactment made possible the identification of two groups of enactments. Two teachers' enactments tended to be a good match for the intended enactment whereas the other two teachers' enactments were less reflective of the intended enactment. Moreover, the distinction between the groups was evident not only in the ratings across analysis categories, but also in the specific aspects of enactments that led to the assigned ratings. In each case, the match of individual teacher's enactment to the respective group was quite reliable.

These groups were also distinguished by students' achievement scores. Effect sizes were statistically significant on high and medium cognitive level questions for students in the first group and were not statistically significant on high cognitive level questions for the second group

(Table 2). Interestingly, only the category of accuracy was not a unique indicator for either group or for student achievement. Teachers who presented science accurately were in both groups.

This analysis identified eight main analysis categories: *accuracy* and *completeness* of science ideas presented, amount student learning *opportunities*, *similarity* of learning opportunities with those intended, and quality of *adaptations*, and amount of *instructional supports* offered, *appropriateness* of instructional supports and *source* of ideas for instructional supports. Rating levels for each category were described (Table 1). These rating levels were effective in discriminating different levels of enactment.

The careful examination of justification statements for patterns in each rating made possible the identification of two to six types of evidence for each main rating category. For example, the types of evidence for *instructional supports* that guided observations and ratings included: 1) *types of instructional supports* –questions, hints and reminders, and real life examples and connections to a driving question, and 2) *activities when instructional support were used* –whole class set up and discussion, small-group work, and student presentations. By rating each of the types of evidence an overall rating for the category was possible and justified.

The identification of eight categories, rating levels, and types of evidence has made it possible to construct scoring rubrics for each category (see Figures 1 – 8). Further, the specific examples used to justify the assigned ratings during the analysis of enactment data described the characteristics of the evidence types that were consistent with high or low ratings for the category in general. For example, under instructional supports when questions are used to guide students to consider important content ideas this evidence contributes to a high rating. Conversely, when questions are used to elicit definition this evidence contributes to low ratings

in for instructional strategies (Figure 6). These descriptions have been added to all rubrics to guide evaluations.

Discussion

Measures used to research teaching in reform need to be reflective of the reform goals. The eight analysis categories are consistent with reform recommendations because they were developed from a reform-based curriculum framework. Moreover, these categories are not specific to the unit used to develop them. Rather the categories should be adaptable to any reform-oriented science program. Any quality program will be concerned about how content is presented, that students have opportunities to learn and that teachers give students guidance and support. The categories were able to separate teachers enactments into two groups that are correspond to two groups indicated by student achievement scores. This suggests the categories and rating levels are capturing something important about teachers' practices that lead to student learning.

The link from specific aspects of classroom teaching to student learning is an important one. Whereas others have shown that teachers effect student learning they have not identified specific instructional practices that lead to improve student outcomes (Darling-Hammond, 1999). In addition, it is an important finding that measuring specific teacher behaviors is not sufficient to determine quality of teaching (Ball & Cohen, 1999; Borko, Cone, Russo, & Shavelson, 1979; Lampert, 1998). It is not enough to know whether teachers are asking questions. It was also important to consider teachers' goals in asking these questions. If we can learn what teachers can do to help students learn we also can learn what types of support teachers need to learn and enact these practices.

The rubrics are based on enactment data and are formatted to facilitate scoring of enactments directly. This should eliminate the need to collect, prepare, and analyze videotape or detail descriptions of classroom events. However, these rubrics have not been field-tested. Although the categories and types of evidence have proven to be useful and informative, other types of evidence may emerge from further observations of reform-based enactments. In addition, we do not know if it is possible to score enactments in real time or if the rubrics will be more usable with videotape that can be paused and rewound. Although much simpler than careful qualitative analysis, these rubrics remain complex. It is likely improvements can be made in rubrics based on use in classrooms or enactment videotape.

The process used to identify categories, rating levels, and specific types of evidence that could be used to characterize teaching was time and labor intensive. However, now that these have been identified future evaluations will be much simpler. Further studies with more teachers enacting reforms would increase the reliability of these recommendations. Through this work, an observation framework that is appropriate for larger scale studies could be created. These categories will be presented in a format easily adapted to various classrooms and curriculum. This will make the much needed large-scale studies of teacher enactments feasible.

We developed these rubrics to evaluate the efficacy of teacher-educative, reform-based curriculum materials but they can be adapted to use in other research questions. For example, Davis (2002) is using student teachers' unit plans to answer questions about how novices learn to teach. Others are using reform-based materials to promote student learning (Prawat & et al., 1992; Songer, Lee, & Kam, 2002). An evaluation scheme like the one presented here would be helpful to gauge how closely enactment reflects the intended curriculum plan without looking for strict implementation (Apple & Jungck, 1990).

This importance of this work lies in its ability to provide a tool to facilitate research on teaching. One area of weakness is the lack of studies that bridge the gap between teacher preparation, classroom teaching, and student outcomes on a large scale. We know that teachers need to learn about teaching in the context of the classroom but we do not know how to efficiently support that learning (Putnam & Borko, 2000). Although we used this observation framework to inform the design of materials to support teachers in reform, this framework will be valuable in many areas of research on teaching. A method to evaluate teaching that is meaningful and usable on a large scale is needed to inform teacher education and professional development research.

References

Anderson, R. D. (2001). The ideal of standards and the reality of schools: Needed research. *Journal of Research in Science Teaching, 38*(1), 3-16.

Apple, M. W., & Jungck, S. (1990). "You don't have to be a teacher to teach this unit": Teaching, technology, and gender in the classroom. *American Educational Research Journal, 27*(2), 227-251.

Ball, D. L. (2000). Bridging practices: Intertwining content and pedagogy in teaching and learning to teach. *Journal of Teacher Education, 51*(3), 241-247.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is--or might be--the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6-8,14.

Ball, D. L., & Cohen, D. K. (1999). *Teacher learning and instructional capacity: Interaction and intervention.* Paper presented at the American Educational Research Association Annual Meeting, Montreal.

Blumenfeld, P. C., Fishman, B. J., Krajcik, J. S., Marx, R. W., & Soloway, E. (2000). Creating usable innovations in systemic reform: Scaling-up technology-embedded project-based science in urban schools. *Educational Psychologist, 35*(3), 149-164.

Blumenfeld, P. C., Krajcik, J. S., Marx, R. W., & Soloway, E. (1994). Lessons learned: A collaborative model for helping teachers learn project-based instruction. *Elementary School Journal, 94*, 539-551.

Borko, H., Cone, R., Russo, N. A., & Shavelson, R. (1979). Teachers' decision making. In P. Peterson & H. J. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 136-160). Berkeley: McCutchan Publishing Corp.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141-178.

Cochran-Smith, M., & Fries, M. K. (2001). Sticks, stones, and ideology: The discourse of reform in teacher education. *Educational Researcher, 30*(8), 3-15.

Collopy, R. (1999). *Teachers use of and learning from curriculum materials.* Unpublished doctoral dissertation, University of Michigan, Ann Arbor.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state policy evidence*: Center for the study of teaching and policy, University of Washington.

Davis, E. A. (2002). Scaffolding prospective elementary teachers in critiquing and refining instructional materials for science. In P. Bell & R. Stevens & T. Satwicz (Eds.), *Keeping Learning Complex: The Proceedings of the Fifth International Conference of theLearning Sciences (ICLS)* (pp. 71-78). Mahwah, NJ: Erlbaum.

Fishman, B., & Best, S. (2000). *Professional development in systemic reform: Using worksessions to foster change among teachers with diverse needs.* Paper presented at the National Association of Research in Science Teaching, New Orleans, LA.

Krajcik, J., Czerniak, C., & Berger, C. (2002). *Teaching science in elementary and middle school classrooms : A project-based approach.* Boston, MA: McGraw-Hill.

Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000). *Inquiry based science supported by technology: Achievement among urban middle school students.* Paper presented at the National Association of Research in Science Teaching, New Orleans, LA.

Lampert, M. (1998). Studying teaching as a thinking practice. In J. Greeno & S. G. Goldman (Eds.), *Thinking practices* (pp. 53-78). Hillsdale, NJ: Lawrence Erlbaum.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks: Sage Publications.

National Commission on Teaching and America's Future. (1996). *What Matter's Most: Teaching for America's Future*. New York: National Commission of Teaching and America's Future.

National Research Council. (1996). *National science education standards*. Washington DC: National Academy Press.

Palincsar, A. S., Magnusson, S. J., Marano, N., Ford, D., & Brown, N. (1998). Designing a community of practice: Principles and practices of the GIsML community. *Teaching and Teacher Education, 14*(1), 5-19.

Prawat, R. S. (1992). Are changes in views about mathematics teaching sufficient? The case of a fifth-grade teacher. *The Elementary School Journal, 93*(2), 195-211.

Prawat, R. S., & et al. (1992). Teaching mathematics for understanding: Case studies of four. *Elementary School Journal, 93*(2), 145-152.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Remillard, J. T. (1999). Curriculum materials in mathematics education reform: A framework for examining teachers' curriculum development. *Curriculum Inquiry, 29*(3), 315-342.

Ruiz-Primo, M. A., Shavelson, R., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 369-393.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessmentsystem (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Schneider, R., Blumenfeld, P., & Krajcik, J. (2002). Designing materials to support teachers in reform. In T. Satwicz (Ed.), *Keeping Learning Complex: The Proceedings of the Fifth International Conference of theLearning Sciences (ICLS)* (pp. 398-405). Mahwah, NJ: Erlbaum.

Schneider, R., & Krajcik, J. (2002). Supporting science teacher learning: The role of educative curriculum materials. *Journal of Science Teacher Education, 13*(2), 167-217.

Schneider, R., Krajcik, J., & Blumenfeld, P. (2002). *Exploring the Role of Curriculum Materials to Support Teachers in Science Education Reform.* Paper presented at the National Association for Research in Science Teaching, New Orleans, LA.

Singer, J., Marx, R. W., Krajcik, J. S., & Clay-Chambers, J. (2000). Constructing extended inquiry projects: Curriculum materials for science education reform. *Educational Psychologist, 35*(3), 164-178.

Songer, N. B., Lee, H.-S., & Kam, R. (2002). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? *Journal of Research in Science Teaching, 39*(2), 128-150.

Wood, T., Cobb, P., & Yackel, E. (1991). Change in teaching mathematics: A case study. *American Educational Research Journal, 28*(3), 587-616.

Table 1 Categories and rating levels of coding scheme used to analyze classroom enactment data

Accuracy

Scientific - all ideas are consistent with current scientific ideas

Sufficient -consistent with current scientific ideas for all main ideas, inaccurate for minor ideas

Semi accurate - inconsistent with current scientific ideas for some main ideas

Non scientific - inconsistent with current scientific ideas for many main ideas

Completeness

Thorough - all the appropriate science ideas are addressed

Sufficient - all the appropriate main ideas are addressed but some minor ideas are missing

Incomplete - missing some main ideas

Insufficient - missing several main ideas

Excessive - includes ideas at a level beyond intended for students

Opportunities

Maximum - includes ample (number or time) opportunity for student learning

Sufficient - includes some (number or time) opportunity for student learning

Insufficient - includes few (number or time) opportunity for student learning

Minimal - includes almost no (number or time) opportunity for student learning

Similarity

High - matched to intended lesson

Medium - closely resembles intended lesson, minor changes

Low - faintly resembles, major changes

None - not consistent with intended lesson

Adaptation

High - adaptation consistent with learning goal and appropriate for students' learning needs

Medium - adaptation consistent with learning goal but <u>not</u> appropriate for students' learning needs

Low - adaptation not consistent with learning goal

None - not adapted

Instructional Supports

High - provides many instructional supports for student thinking

Medium - provides some instructional supports for student thinking

Low - provides few instructional supports for student thinking

None - provides no instructional supports for student thinking

Appropriateness

Excellent - instructional supports always used in ways matched to student learning needs

Sufficient - instructional supports usually used in ways matched to student learning needs

Insufficient - instructional supports usually not used in ways matched to student learning needs

Poor - instructional supports always used in ways not matched to student learning needs

Sources

Supplemented - used instructional supports included in materials plus others

Matched - used only instructional supports included in the materials

Replaced -used only instructional supports not included in materials

Table 2
Student performance on pre- and post-tests for each teacher.

	Pre-test M (SD)	Post-test M (SD)	Effect Size[a]
Enactment Group One			
Ms Franklin, Fall 1998 (N = 29)			
High level (18 points)	1.66 (1.08)	3.97 (2.23)	2.14***
Medium level (19 points)	6.34 (1.45)	10.03 (2.23)	2.54***
Low level (16 points)	8.03 (2.28)	9.59 (3.21)	0.68*
Overall (53 points)	16.03 (3.45)	23.59 (6.16)	2.19***
Ms Wells, Fall 1999 (N = 56)			
High level (4 points)	0.63 (1.59)	1.25 (1.96)	1.06***
Medium level (9 points)	3.79 (1.39)	4.41 (1.69)	0.45*
Low level (8 points)	2.73 (1.27)	3.63 (1.36)	0.70***
Overall (21 points)	7.14 (2.11)	9.29 (3.04)	1.01***
Enactment Group Two			
Mr. Davis, Fall 1999 (N = 25)			
High level (4 points)	0.44 (0.65)	0.72 (1.02)	0.43
Medium level (9 points)	3.60 (1.32)	3.72 (1.51)	0.09
Low level (8 points)	2.40 (1.29)	4.40 (1.85)	1.55***
Overall (21 points)	6.44 (1.87)	8.84 (3.16)	1.28***
Ms Turner, Fall 1998, (N = 25)			
High level (18 points)	0.88 (1.05)	0.88 (1.01)	0.00
Medium level (19 points)	5.04 (1.90)	6.00 (2.40)	0.51*
Low level (16 points)	4.48 (2.22)	6.00 (2.87)	0.68*
Overall (53 points)	10.40 (3.31)	12.88 (5.10)	0.75**

[a]Effect Size: effect size was calculated by the difference between the means divided by the standard deviation of the pre-test.

*$p < .05$. **$p < .01$. ***$p < .001$.

***Accuracy* of science ideas presented**

Types of evidence	***Non scientific*** Inconsistent with current scientific ideas for many main ideas	***Semi accurate*** Inconsistent with current scientific ideas for some main ideas	***Sufficient*** Consistent with current scientific ideas for all main ideas but inaccurate for minor ideas	***Scientific*** All ideas are consistent with current scientific ideas
Explicit statements Definitions				
Explanations				
Examples				
Guidance Direction of student attention to tasks	*Completion of tasks* *Irrelevant factors*			*Conceptually important aspects* *Appropriate ideas*
Guidance in connection with student predictions, hypothesis or conclusions	*Little guidance* *Guided students to inappropriate form of statements*			*Guided students to appropriate form statements*
Guidance in connection with student investigation design	*Little guidance* *Guided students to incomplete or inappropriate design*			*Guided students to complete and appropriate design*
Response to students Accurate and inaccurate student statements	*Not distinguished*			*Distinguished; inaccurate redirected, accurate acknowledged*
Inaccurate student statements during presentations	*Generally not corrected*			*Corrected*
Overall rating				

Figure 1: Types of evidence and rating levels for the category of Accuracy

Completeness of science ideas presented

Types of evidence	***Insufficient*** Missing several main ideas	***Incomplete*** Missing some main ideas	***Sufficient*** All the appropriate main ideas are addressed but some minor ideas are missing	***Thorough*** All the appropriate science ideas are addressed
Intended content	*Not addressed or only defined*			*Addressed*
Concepts intended for the lesson sequence				
Process ideas regarding investigations (variables and design)				
Process ideas regarding graph reading and interpretation				
Generalizable statements				
Connections between ideas	*Not explicit or not made*			*Explicitly addressed*
	Excessive Includes ideas at a level beyond intended for students			
Outside content	*Yes*			*No*
Content beyond that intended				
Overall rating				

Figure 2: Types of evidence and rating levels for the category of Completeness

Amount of Student Learning *Opportunities*

Types of evidence	***Minimal*** Includes almost no (number or time) opportunity for student learning	***Insufficient*** Includes few (number or time) opportunity for student learning	***Sufficient*** Includes some (number or time) opportunity for student learning	***Maximum*** Includes ample (number or time) opportunity for student learning
Time	*Class time short for all activities except final student presentations*			*adequate in class time for each type of activity*
Type of activity Actions	*Incomplete*			*completed*
Small-group work	*Limited* *Includes little thoughtful work*			*Frequent* *Included action and thoughtful work*
Discussion	*Limited* *Few student ideas used*			*Frequent* *Used student ideas*
Structure Activities	*Clustered by type*			*Sequenced and cycled*
Small-group work	*Monitored closely for completion*			*Monitored but not overly structured* *Students allowed to discuss and work together*
Discussions	*Presented teacher ideas and explanations*			*Used student ideas* *Either clearly focused and directed <u>or</u> followed student ideas*
Investigations	*Structured by list of items to complete* *Not structured*			*Structured by question*
Overall rating				

Figure 3: Types of evidence and rating levels for the category of Opportunities

Similarity of Student Learning Opportunities

Types of evidence	***None*** Not consistent with intended lesson	***Low*** Faintly resembles, major changes	***Medium*** Closely resembles intended lesson, minor changes	***High*** Matched to intended lesson
Major learning opportunities Overall opportunities				
Phases of opportunities				
Sequence Of overall opportunities				
Of phases of opportunities	*Combines like activities*			
Emphasis				
Overall rating				

Figure 4: Types of evidence and rating levels for the category of Similarity

Adaptation of Student Learning Opportunities

<table>
<tr><th>Types of evidence</th><th>None
not adapted</th><th>Low
adaptation not consistent with learning goal</th><th>Medium
adaptation consistent with learning goal but <u>not</u> appropriate for students' learning needs</th><th>High
Adaptation consistent with learning goal and appropriate for students' learning needs</th></tr>
<tr><td rowspan="2">Additions</td><td colspan="4">Group presentations
Non-content supporting features
Does not adapt
More whole class activities to address students' questions
Investigation features such as variables
Final presentation features such as questions or demonstrations of design</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>Changes</td><td colspan="4">Teacher-led activities changed to student activities
Small-group activities changed to individual work</td></tr>
<tr><td>Overall rating</td><td></td><td></td><td></td><td></td></tr>
</table>

Figure 5: Types of evidence and rating levels for the category of Adaptations

Amount of *Instructional Supports*

Types of evidence	***None*** Provides no instructional supports for student thinking	***Low*** provides few instructional supports for student thinking	***Medium*** provides some instructional supports for student thinking	***High*** provides many instructional supports for student thinking
Types Questions	*Used to elicit definitions <u>or</u> sometimes explanations*			*Used to guide students to important content ideas*
Hints and reminders	*Used as lists of items to complete*			*Used to focus attention on content related aspects of activity and to guide doing a task*
Real life examples and connections to driving question	*Rarely or occasionally used*			*Frequent*
Activities Whole class set-up and discussion	*Few supports; tasks may be student self-guided work*			
Small-group work	*Frequent prompts to complete*			*Few interruptions*
Presentations	*Few supports*			*Guiding questions*
Overall rating				

Figure 6: Types of evidence and rating levels for the category of Instructional Supports

***Appropriateness* of Instructional Supports**

Types of evidence	***Poor*** Instructional supports always used in ways not matched to student learning needs	***Insufficient*** Instructional supports usually not used in ways matched to student learning needs	***Sufficient*** Instructional supports usually used in ways matched to student learning needs	***Excellent*** Instructional supports always used in ways matched to student learning needs
Questions and prompts	*Answered or explained by the teacher* *Guide students to definitions or voting on right answers*			*Guide students to focus on appropriate ideas*
Hints and reminders	*Address task completion*			*Address ideas with which students may have trouble*
Students ideas	*Not requested*			*Requested* *Connected to previously stated students' ideas*
Feedback	*Identifies mistakes or wrong answers*		*Directs students to appropriate ideas*	
Student questions and difficulties	*Not addressed*			*Addressed*
Overall rating				

Figure 7: Types of evidence and rating levels for the category of Appropriateness

Sources of Instructional Supports

Types of evidence	***Replaced*** Used only instructional supports not included in materials	***Matched*** Used only instructional supports included in the materials	***Supplemented*** Used instructional supports included in materials plus others
From materials	*Many suggested supports not used*		*Uses questions to guide discussion* *Uses driving question* *Comparisons to similar previous activities* *Monitored groups*
Teacher added	*None or* *Prompts for task completion and definitions*		*Real-life examples* *Supports from earlier parts of the materials to later lesson sequences*
Trend	*Matches throughout or matches early then quickly replaces*		*Matches early, but quickly supplements*
Overall rating			

Figure 8: Types of evidence and rating levels for the category of Sources

SE067680

ERIC

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Observing teaching: A reform-based framework for looking into classroom

Author(s): ~~Schneider~~ Rebecca Schneider, Phyllis Blumenfeld

Corporate Source: National Association for Research in Science Teaching Annual meeting

Publication Date: March 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Rebecca Schneider	Printed Name/Position/Title: Rebecca Schneider
Organization/Address: University of Toledo	Telephone: 419-530-2504 / FAX: 419-530-8459
	E-Mail Address: Rebecca.Schneider@utoledo.edu / Date: 3-25-03